DOCUMENT RESUME

ED 172 425 EA 011 857

AUTHOR Jovick, Thomas D.

TITLE Approaches to Data Analysis in Longitudinal Field Investigations of Educational Programs.

INSTITUTION Oregon Univ., Eugene. Center for Educational Policy and Management.

SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE May 79
NOTE 29p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; Data Analysis; Elementary Education; Field Studies; *Longitudinal Studies; Multiple Regression Analysis; Multiunit Schools; Raw Scores; *Research Methodology; True Scores
IDENTIFIERS Management Implications of Team Teaching Project

ABSTRACT

The federally funded longitudinal field study called Management Implications of Team Teaching (MITT) required a search for an appropriate strategy for analyzing through-time relationships among selected variables. The MITT project used questionnaires and interviews to collect data concerning the work, governance, attitudes, and orientation of teachers in fourteen multiunit elementary schools and thirteen conventionally organized elementary schools. In a study like MITT, one typically wishes to investigate gains or decrements in scores on variables as a consequence of either treatment differences (multiunit as opposed to conventional organization) or variation in quantitative variables. Two problems confront investigators who embark on longitudinal data analyses: one concerns whether to use original scores or "true score" estimates, and the other concerns what statistical procedures to apply to these scores. This document explains why and how the investigators in this case chose to use original scores and multiple linear regression analysis. While data from the MITT project are used to illustrate the process by which these methods were chosen, the findings of the original study are not presented here. (Author/PGD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED172425

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

APPROACHES TO DATA ANALYSIS IN LONGITUDINAL FIELD INVESTIGATIONS OF EDUCATIONAL PROGRAMS

By

Thomas D. Jovick

Center for Educational Policy and Management

University of Oregon

Eugene, Oregon

May 1979

The research reported herein was supported in part by funds from the National Institute of Education, U.S. Department of Health, Education, and Welfare. Opinions expressed in this report do not necessarily represent the policies or positions of the National Institute of Education, nor should any official endorsement of the report be inferred from the reporting agencies.

EA 011 857

This report originated in the search for an appropriate strategy for analyzing through-time relationships among selected variables in a federally funded longitudinal field study called Management Implications of Team Teaching (MITT). It explains the rationale for using observed scores rather than true score estimates as the relevant data for analysis and describes the use of multiple linear regression as the primary method of statistical analysis of through-time relationships. Much of the discussion will allude to the MITT Project for concrete examples of longitudinal data analysis problems and characteristics of relevance to many field research studies in education. The description below outlines the dimensions of the MITT study.

The MITT Study

The Management Implications of Team Teaching Project collected data concerning the work, governance, attitudes, and orientations of teachers in a group of multiunit elementary schools and a matched set of conventionally organized elementary schools. Questionnaires and interviews were administered to teachers in 27 schools every six months ending in the Spring of 1976. Fourteen were innovative schools which had installed a multiunit organization of their faculties in the Fall of 1974, two years prior to the final data collection.

The shape of the formal managerial structure followed the Multiunit School model that had been developed by Klausmeir and others at the Wisconsin Research and Development Center for Cognitive Learning. Teachers in each of the 14 multiunit schools organized into two or more teams or units and as a

group held responsibility for the instructional program of students assigned to them. Each unit had an official leader who sat with the principal and other team leaders on a cabinet.

Project staff selected the 13 other schools from the same districts as the multiunit schools to act as matched controls. None of these had adopted any such structural alteration nor equivalents during the course of the MITT investigation. Their managerial structures reflected the undifferentiation typical of the conventional elementary school with the instructional staff responsible for students in their own classrooms at one level and the principal at another.

The criteria for distinguishing the two sets of schools lay solely in the formal organizational structure--presence or absence of teaching units having group instructional responsibility. Behavioral differences constituted an outcome of rather than a distinguishing factor between the schools, e.g., the actual incidence of meetings among team members, the acceptance of collective responsibility on students' instruction, or the actions of leaders did not enter into the multiunit vs. conventional distinction.

Other structural and staff background characteristics accompanied the multiunit innovation and added to the distinction between the two sets of schools. With the innovation came a bundle of smaller innovations most of which were optional as instructional and curricular reforms and aids including IGE instructional materials from Wisconsin. No such innovations appeared in the nonunitized schools although teachers here and there occasionally tried novel methods and materials. The multiunit schools generally were staffed by younger and less experienced teachers more progressive in educational outlooks than were the conventional schools.

All schools in the sample lay east of the Mississippi River, primarily in the New England, Mid-Atlantic and Border States. They included inner-city schools in low income areas, suburb schools for the affluent and for the not-so-affluent, and rural schools; most were public but some were Roman Catholic parish schools; the grade organization of some included seventh and eighth grades, others were exclusively primary; size ranged from an eight teacher faculty with 200 students to a 30 teacher faculty with over 800 students.

More detailed reports concerning conceptual development, definition, and operationalization of variables and findings can be found elsewhere (Packard, et al., 1978; Packard, et al., 1976; Charters, 1978; Packard and Jovick, 1978; Duckworth and Jovick, 1978; Packard, 1975; Charters, 1975a; Packard and Carlson, 1975: Jovick and Milich, 1975; Charters, 1975b).

Rather than refer to initial (pre) and final (post) variation on variables in designs with a series of measurement points, it is easier to refer to each point as T1, T2, etc., where T1 is the first data collection, T2 the second, and so on. For the MITT study, T1 is the pre-unitization data collection representing initial variation and T2 through T5 are post-unitization measures with T5 being considered final variation.

Preliminary Considerations

The difficulties in establishing a through-time individual-level file demanded that the majority of MITT's early analyses use only school-level indices, thereby limiting the effective sample size to the number of schools. This subsequently influenced our perspective on what kinds of scores to employ in our analysis of change and with what statistical procedures to analyse them.

A number of logistical difficulties typically preclude the ready creation of a through-time individual-level file in longitudinal field studies. These constraints require that the majority of the early analyses use data at a higher, more easily compiled unit of analyses. A decision to limit analyses to a higher level which would require the use of aggregated data reduces the effective sample size and influences the perspective on what kinds of scores to employ in the analysis of change and with what statistical procedures to analyze them. Such was the case with the MITT data for which the enormous problems of individual-level longitudinal file construction and the requirements to meet report deadlines demanded that we conduct the majority of our analyses at the school level, thereby limiting our usable sample size to the number of schools.

A researcher may frame more substantive reasons for employing a higher level of analysis like the school. In MITT, we had conceptualized some major variables as properties of the school and expected they would change over time in response to the anticipated installation of the multiunit structural organization among the faculty. Although many of our major interests focused on the change in beliefs and activities of individual teachers, a school level analysis was reasonable because the implementation of the innovation was to be a school-wide phenomenon affecting the behavior of most, if not all, teachers in the school.

Furthermore, all 27 schools participated throughout the course of the study, even though the teacher turnover gave us slightly different staff compositions at each wave. In fact, only about two-thirds of the faculty in all multiunit and conventional schools at T1 were present in the schools

at T5. By using the school as the unit of analysis we did not have to confine our information about variables to that given by this two-thirds faculty cohort.

This does not imply we eliminated the individual teacher as a unit of analysis; many of our later cross-sectional and longitudinal analysis used the teacher as the unit (Packard, et al., 1976; Packard, et al., 1978). This was especially true for the teacher attribute and perceptual variables which conceptually characterize properties of individuals rather than organizations. For the school-level analyses, we aggregated many of these to depict mean levels of selected teacher attributes in each school (Packard, et al, 1976, 1978).

In a study like MITT, one typically wishes to investigate gains or decrements in scores (or their aggregated counterparts) on variables as a consequence of either treatment differences (multiunit vs. conventional organization) or variation in quantitative variables.

Two problems confront investigators who embark on longitudinal data analyses: one concerns whether to use original scores or "true score" estimates, the other concerns what statistical procedures one should apply to these scores.

Strategies for Estimating True Scores

Among the alternate approaches to constructing scores which reflect change have generally been those like the "raw difference" or "raw gain" scores formed by subtracting Tn-1 scores on a variable from its Tn

scores, or "residual gain" scores formed as deviations of Tn scores from predicted values based on a regression of Tn scores on Tn-1 scores. (Note that these difference and residual gain scores can be constructed using larger lags such as that from Tn-3 to Tn; this depends upon the relationship of interest.) Cronbach and Furby (1970) quite emphatically claim that statistical analyses should not use these types of scores since they will yield either the same results as an analysis on the original scores at Tn or results more difficult to interpret.

The basic thrust of the "true score advocates" like Cronbach and Furby and the others discussed here is that, although original scores on variables are better for statistical purposes than the typical corrections for initial scores, they too leave much to be desired precisely because they contain error.

Their strategies encompass the same general approach--to use all pertinent information to control for sources of measurement error and thereby derive more accurate estimates of true scores. Essentially their proposals describe measurement models for relating true scores to their observed values and rely upon assumptions of classical test theory. The resulting true score estimates become the basis for subsequent analysis of relationships of final status to treatment conditions or to variation in other variables.

Classical test theory proponents conceptualize any observed score as a combination of true score plus measurement error. A "true score" for an individual (or the analysis unit of interest) is the average score over a large number of repeated measurements of a variable at a particular time

point. Factors which do not relate conceptually in a systematic way to the characteristics one wishes to observe or the inferences one wishes to examine are sources of measurement error in scores (Thorndike, 1971, Ch. 13) and account for part of the variation among observed scores at any particular data collection. Over a large number of repeated measurements of a variable at a particular time, the average of the measurement errors equals zero.

A piece of information crucial to the determination of "true" scores is the instrument's reliability, an index of the amount of error in one's measurement. The variability of the sample under study affects the reliability estimate--a reason for requiring large random samples in reliability studies. One way of expressing the coefficient is as follows:

$$r_{XX} = 1 - \frac{\sigma^2_{meas}}{\sigma^2_X}$$

The variance of the errors of measurement (σ^2_{meas}) is considered to be approximately independent of the variance of the obtained scores (σ^2_X) and is therefore conceptually regarded as a fixed characteristic of the instrument regardless of the sample being studied. As the variance of the sample increases then, the ratio $\frac{\sigma^2_{meas}}{y^2}$ will decrease and the reliability r_{XX} will increase; as the variance decreases the ratio will increase and the reliability decrease.*

*Nunally (1967) notes that a low reliability for an instrument will make detection of statistical significance difficult; when the standard error of measurement and the standard deviation of the variable in the sample are approximately equal he claims it is hopeless to investigate the variable.

The problem for analysis is that measurement errors at any point in time may distort the assessment of relationships among variables if the analysis does not explicitly account for them. Longitudinal designs compound the problem because a portion of a measure of relationship between variables at any two points in time includes a correlation between measurement errors that occur at each time.

Correction for Attenuation

This is the simplest and most straightforward strategy for estimating true scores or relationships between them. There are two approaches.

One involves calculating the correlation between two variables that would occur if they were both perfectly reliable. It entails using the reliability coefficients in the following formula:

$$\bar{r}_{12} = \frac{r_{12}}{\sqrt{r_{11} r_{22}}} = \text{corrected } r_{12}$$

The second approach involves obtaining estimates of unbiased scores because obtained scores tend to be biased, i.e., high scores tend to be higher than their true score counterparts and low scores tend to be lower. Conceptually, unbiased scores are those that people/schools/units would obtain if they were administered all possible tests having equal numbers of items sampled randomly from the same domain--they are estimates of true scores. In the formula below,

$$t' = r_{XX} x$$

r_{XX} is the reliability estimate for the variable X, $x = (X - \bar{X})$ and t' is

a true score estimate in deviation score units. By adding t' to $\overline{X}$ one obtains an estimated true score.

Nunnally recommends that estimating true scores is necessary only in longitudinal analyses where one is interested in contrasting the changes between groups. To correct obtained scores O'Connor (1972) advocates using the test-retest reliability; if it were unavailable then some measure of internal consistency would be satisfactory.

Cronbach and Furby Method

These authors extend the idea of correcting raw scores to studies involving two or more groups. Their strategy is to estimate true scores for independent and dependent variables within treatment and control groups and then to enter these true scores into regression equations. Their calculations for a true score on any variable employs more information than the instrument's reliability coefficient alone.

Let X1 and X2 represent time 1 (T1) and time 2 (T2) scores on variable X. Unless the true correlation between the two is zero, both X1 and X2 contain information about the true score for X1, here indicated as X1t, and both can be used in a regression equation to obtain predicted true X1t scores.

The reliability coefficient provides information about the X1t's on the basis of the actual X1 scores. The deviations of X2 values from those predicted by the regression of X2 on X1 within the experimental and control groups separately furnish information about the X1t's on the basis of X2 scores.

They present the following general formula (p. 71), expressed here in terms of X1 and X2:

$$\hat{X}1t = r_{XX}X1 \quad + \quad B_{X1t(X1 \bullet X2)}\ (X1-X2) \quad + \quad (1 - r_{XX})\ \overline{X1}$$

Where X1•X2 is their notation to represent residual scores formed from predicting X2 from X1 and the final expression is an adjustment of the mean of the group in terms of the lack of correlation between true and obtained scores, i.e., unreliability. Presumably, the reliability coefficient used in the equation would also be calculated within each group, although the authors do not specifically say so.

One can pool the groups to obtain single "within-group" values for the parameters in the above equation but the estimates will be better when calculated separately within groups. This is especially true for groups not formed randomly because the true score distributions within each tend not to be the same; this implies that the same observed score, say X1, has a different true score, X1t, depending upon the group.

Cronbach and Furby go on to say that one should include other relevant T1 variables in the true score estimate, the major limitation being the sample size used in the regressions. Wiley and Harnischfeger (1973) qualify that recommendation advocating the use of such other variables only if one can defend, through a causal model, that they are theoretically direct determinants of the X2 variable. Their impression is that Cronbach and Furby would throw a large number of T1 background and other variables into the analysis indiscriminately in the hopes of reducing error.

Cronbach and Furby provide a method for calculating true score variances and then inputing these into regressions rather than using raw scores. They also distinguish between linked and unlinked T1-T2 measures and furnish adjustments that need to be made respectively.

Wiley and Hornik Method

Wiley and Hornik (1973), to get accurate and true score estimates, developed a measurement model to deal with errors in panel data but it relies on obtaining more than one measure of each variable at each point in time. The two measures for any single variable reflect the same true score variable but different error variances. Armed with assumptions from classical test theory about the independence and additivity of variance components, they use the cross-time and alternative observed scores to estimate the unobserved true and error variance in the variables. The true variances and covariances of the variables serve as the basis for computing regression weights for the relationships among variables over time.

Hannon and Others

Wiley and Hornik allude to more "optimal" methods for using multiwave longitudinal data, more optimal in the sense that they reduce the standard error around the estimates. The necessary calculations use the pieces of variance information that are available in all data collection points. The references they cite, particularly for the relevant computer program, suggest these methods are variations of confirmatory factor analysis.

John Meyer* suggested using confirmatory factor analysis only if one found statistical significance with a technique developed by Michael Hannon and Alice Young (ND). Their method attempts to pool variance information about the variables of interest across all waves in an effort to get a more accurate handle on statistical significance; like others, the strategy focuses upon reducing the amount of error variance to get an accurate estimate of true score relationships. The Hannon-Young method and confirmatory factor analysis require complicated estimation procedures which cannot be performed with least squares regression analyses.

Problems With the Strategies

Some general controversy exists over the use of corrected correlations or true score estimates in analyses:

(1) One may be fooling oneself into believing that a better correlation has been uncovered than the one actually obtained. Nunnally (1967) notes that correlations corrected for attenuation seldomly differ much in magnitude from the obtained correlations.

(2) The correction itself may be poor if the reliability estimates themselves are poor. Moreover, since the possibility exists for reversals in signs for regression coefficients and partial and part correlations if one uses corrected rather than actual correlations, he must have confidence in the reliabilities in order to have confidence in the regression output from corrected correlations.

*Personal communication, Stanford University

(3) Nunnally adds that in prediction type problems it may be inappropriate to correct for unreliability in the criterion since the issue is to predict or explain scores on that variable as they actually exist not as they would exist were the test perfectly reliable. (He seems to be particularly addressing prediction problems related to selection decisions.)

We seriously questioned the appropriateness of the Hannon-Young and Confirmatory Factor Analysis techniques for analysis in field investigations like MITT. Few studies have subjected their model to critical examination. Although their single report appears to lend support to its utility, the circumstances for its applicability remain clouded in a number of assumptions about the characteristics of the data one wishes to analyze. Similarly, the application of the Wiley-Hornik method, although simple, also appears to rest on assumptions which do not characterize the nature of most longitudinal research data in education. Of any of the methods, the Cronbach-Furby approach and the Correction for Attentuation seem the most straightforward. Yet, four major points leave them open to question also.

(1) The Cronbach-Furby and the Attentuation Correction methods have their focus on research using standardized tests of abilities, I.Q., and personality traits which themselves have a tradition of being extensively researched for reliability and validity on large random samples of subjects. Consequently, the reliabilities of such measures are generally acknowledged as being stable and accurate and not affected by the variance of research samples in which they are used. We strongly doubted the suitability of their formulas for many of the kinds of information collected in field investigations.

For example, MITT did not collect data about abilities, I.Q., or personality traits. It addressed different phenomena such as opinions about others, perceptions about the school, perceptions of self, and descriptions of behavior. None of the instruments received the extensive attention that trait and ability measures have traditionally received nor were their reliabilities calculated on exceedingly large random samples. This is not to say they are no good; it just questions the amount of confidence we can place in the accuracy of the reliabilities that would be used to estimate true scores. Questionable reliabilities would guarantee us nothing better than questionable true score estimates.

(2) The accuracy of the reliabilities for true score computation relates to a more difficult problem: how should one compute reliability in a study like MITT?

(a) Should one calculate it separately for each "treatment" group (multiunit vs. conventional), then differences in the variance of each group will produce different estimates even though a reliability, like the standard error of measurement, is conceptually a characteristic of an instrument independent of any sample of subjects.

(b) Can one justifiably use reliabilities based on individuals when the unit of analysis is the school? This is part of an aggregation problem discussed by Hannan (1971) which maintains that aggregated scores measure a theoretically different variable than the unaggregated scores; perhaps reliabilities should be based on school scores rather than individual scores.

(3) The Cronbach-Furby method includes a term consisting of residuals, deviations of X2 from predicted X1 scores as calculated within each "treatment" group. The logic of using post-test to predict pre-test is that:

> Within a treatment, persons higher on the post-test than others having the same observed pre-test score tend to be those for whom the true pre-test score is higher than the observed score. (p. 72)

This may apply to ability and trait measures but it may not accurately model many of the phenomena under study in field investigations. For example, in MITT we may find that an experimental school has more classroom communication among teachers than another experimental school but that both have the same amount of communication at the previous wave. The contention of Cronbach and Furby is that the first school should theoretically have a higher "true" communication level at the previous wave also. Such logic left us doubting further their method's utility.

(4) Finally, whether one employs variances of raw score regressions to compute true scores according to the Cronbach-Furby approach, he still faces the prospect of error in the variances due to small sample size. This is especially the case if true scores for variables must be estimated within each group.

In summary, we found no guarantee that any of the methods above would provide us accurate and, therefore, useful estimates of true scores. The arguments for use of the methods made sense but we remained unconvinced how accurately the models for handling error reflect the characteristics of our design and variables. (We do believe a study to employ the models on the MITT data is something that could be worthwhile in itself.) Consequently, we

decided not to estimate true scores, but to proceed in our longitudinal analyses with the uncorrected indices.

Selecting the Statistical Method for Longitudinal Analysis

Our preliminary queries about across-time analyses initially centered upon the detection of experimental-control differences at the various points in time while taking into account temporal differences in variation in the dependent variable of interest. Our naivete led us to attempt to fit the Repeated- Measures ANOVA to our design but we eventually deemed that model wholly inappropriate.

Although approaches in the analysis of time series experiments also seemed relevant, initial explorations suggested the MITT study lacked a sufficient number of data points both before and after unitization for the technique to be useful. No more will be said of this approach other than that the possibility for some form of time series analysis may be applicable; the problem lies in using the conceptual and mathematical properties to identify the model which accurately reflects the appropriate moving average, autoregressive, and stationary characteristics of the MITT design and variables (Glass, Willson and Gottman, 1975).

We settled upon the generalized multiple linear regression. The approach relates the status of a dependent variable at one point in time to the status of the same and/or other variables at previous points in time. Its use in analysis of covariance is quite amenable to the study of gain or loss (change) as a function of treatment (Pelz and Lew, 1970). In addition, it permits the assessment of curvilinear and contingent/interactive

relationships (Cohen and Cohen, 1975; Amick and Walberg, 1975). In some respects it is a residual gain strategy criticized by Cronbach and Furby. The impossibility of random assignment of schools to groups (multiunit or conventional) meant we had no highly rigorous method at our disposal to control for differences in the groups prior to the implementation of the innovation. At best, the conventional schools constituted a set of control schools matched with the multiunit schools on district and state characteristics (Hummel-Rossi and Weinberg, 1973).

We used three types of information from most regression analyses. One is the total proportion of variance in Y that is accounted for by the independent variables together. Another is the increment in the proportion of variance explained by the addition of a variable to the prediction equation. A third is the regression coefficient for each variable, usually beta weights, to indicate its unique effect upon the dependent variable.

Because we conducted our analyses at the school level, the restricted sample size limited the number of independent variables we could use in any single prediction equation. We usually limited the number to two and often computed an additional interaction variable to include as a test for the homogeneity of regression. Had we found significant interaction we would examine separate regression lines which the analysis indicated were different. If no significant interaction were found, we would report regression coefficients from the equation with the interaction term omitted (Kerlinger and Pedhazur, 1973).

Repeated Measures ANOVA and Longitudinal Field Data

Historically, laboratory researchers have found repeated measures designs useful for studying the same unit of observation under several different treatments, conditions, or trials. For example, several different ways of learning a certain type of material or several trials of rehearsal on a list of words comprised the repeated treatment for each subject. In these cases, the researcher randomized the order of presentation of the treatments (e.g., of the words in the list).

The design usually characterizes experiments concerned with accounting for effects due to learning, transfer, and fatigue. A main effect for the repeated dimension indicates that, regardless of the order in which subject received the treatments or conditions, their scores consistently increased or decreased over time. One typically interprets an increase in terms of rehearsal/practice/transfer of learning concepts, a decrease in terms of fatigue/motivational concepts.

The pre-post design common to educational field research and evaluation does not fit this paradigm. The levels of the repeated dimension do not coincide with the administration of a treatment or condition; the treatment, rather, intervenes between a pair of points in time.

If MITT had used the repeated measures ANOVA, any main effect found for the repeated factor would pose interpretive problems due to the lack of correspondence in the design models; although it would indicate that scores changed (increased or decreased) over time, no useful concepts like practice, transfer of learning, fatigue or motivation exist to which we could reasonably attribute that change.

We could not unambiguously ascribe the effect to unitization since the repeated dimension main effect reflects a through-time averaging over both unitized and nonunitized schools. If anything, we would expect to find such across-time effects due to the multiunit-conventional distinction showing up as a significant interaction; however, the pre-post design for the research and the nature of the "treatment" in a study like MITT guarantee a time-by-treatment interaction.

Moreover, the use of the repeated measures model to analyze pre-post or any time-to-time data violates assumptions crucial to a repeated measures paradigm. One of these is that the correlations among the levels of the repeated dimension (trials, treatments, conditions) are the same. With random ordering of the trials or treatment conditions for each subject, one need not show concern about the carry-over effects from one trial to another. The covariation among trials is distributed throughout the sample and, as one of the powerful characteristics of the analysis, is accounted for by partitioning it out of the within-subject variation. In the pre-post type of field study design, the order of the data collection waves cannot be randomized among the schools; any covariation among them will therefore be nonrandom.

The repeated measures ANOVA falls short also because one of its distinctive characteristics lies in its great sensitivity to detecting within-subject effects but relative insensitivity to detecting between-subject or, in the MITT case, multiunit-conventional effects. The covariation among trials represents explained variance due to variation within each subject

(school) and the computational procedure extracts it, thereby reducing the unexplained or residual portion of variance; consequently, the analysis reduces the size of the residual term by an amount determined by the degree of covariation among the trials. This smaller residual term, as a divisor for the F-statistic used to test within-subject effects, is smaller than it would be had there been no repeated dimension and will increase the chances of finding a significantly large F value.

Application of Lagged Multiple Linear Regression on Longitudinal Field Data

MITT used regression analyses to address two general types of goals: one was the detection of T5 differences between multiunit and conventional schools, the other was the assessment of lagged relationships among variables.

Impact of the Innovation

In assessing the impact of a treatment or innovation over time, one often finds the groups differ on several dependent variables of interest prior to treatment or implementation of the innovation. The accurate statistical determination of differences between or among groups must then take these prior differences into account. Fortunately, the availability of such data about the dependent variables affords one an opportunity for a more powerful analysis. In MITT, the multiunit and conventional schools differed on a number of crucial variables prior to the formation of teaching units. Consequently, we could not confidently attribute differences detected at T5 to the effects of unitization unless we controlled for the T1 differences.

To do this we employed the multiple regression approach by regressing T5 values of a dependent variable first on the T1 values (auto regression variable) of the same variable, and then on a second variable, a dummy-coded School Organization variable coded with 1's for multiunit and 0's for conventional schools.

Since the procedure is the regression application of the analysis of covariance, our interest was in the increment in the proportion of variance accounted for by School Organization beyond that explained by the T1-T5 autocorrelation (Kerlinger and Pedhazur, 1973). In the same regression equation we checked for an interaction effect (homogeneity of regression) to determine if the influence of unit organization on T5 values were contingent upon T1 values.

At times, it is wise to exclude the pre-innovation autoregression variable when assessing the impact of the innovation or treatment on particular dependent variables. These are cases in which high colinearity exists between the autoregression variable and the treatment variable of interest. If the autoregression variable correlates strongly with the independent variable which reflects differences or variations in treatments or innovations (as did the MITT School Organization variable), then the regression estimates cannot clearly separate out effects of the treatment or innovations. In such cases, a substantial amount of variance shared by the autoregression variable and the independent variable may produce nonsignificant and low effect coefficients for both (Gordon, 1968). Because the regression procedure would "remove" the effect of interest, the interpretation that no relationship existed could be erroneous.

This may happen, as it did in the multiunit schools, in schools where teachers "gear up" for an innovation through pre-implementation participation in activities and display of attitudes that the innovation alledgedly will foster once installed. By the time it is implemented the school faculty may already have been doing some of what the innovation was to induce them to do.

Lagged Relationships Among Variables. One also frequently wishes changes in outcone variables as a function of other outcome or nonprogram variables. Changes in some outcome variables may result directly from "treatment" or innovation and, in turn, induce changes in other variables. In using regression analysis for this purpoes one must assume that the length of the lag used in the analysis accurately depicts the actual delay in the influence of the variation of the earlier variable on that of the later variable.* For this purpose we usually assumed that the most relevant through-time influence on a dependent variable came from variation in the immediately previous wave, six months prior. The regression analysis attempted to find evidence that the Tn+1 variation in a dependent variable was influenced by variation in other variables that occurred at Tn.

---

* Heise (1970) did some early work on this approach for variables at two points in time, and Pelz and Lew (1970) extended his model for analysis to cover multiple waves of data. Pelz and Lew used a Monte Carlo approach to study the relationship of length of lag to accuracy of regression estimates under varying conditions of true lags. Because it addressed a system of relationships in equilibrium, uninterrupted through time by planned change, their formulation and results have limited usefulness for the processes examined in field studies of planned innovations.

Procedurally, Tn+1 values of a dependent variable, Yn+1, are regressed on Tn values of the same variable, Yn, and the Tn values of a different variable, Xn, which is hypothesized to influence it. Yn is called an auto-regression variable and controls for stability in the variation of Y across the lag. The effects attributed to Xn, then, represent its incremental influence over Y over the lag after controlling for whatever relationship it had to Y at Tn.

The Tn-Tn+1 autocorrelation of the dependent variable of interest reflects the extent to which its variation at a particular wave, Tn+1, came about in response to or at least as some predictable function of its variation at immediately previous wave. A low autocorrelation would suggest that its variation at Tn+1 is a function of something other than its variation at Tn; a high autocorrelation would suggest its variation came about at least in partial response to its previous variation and perhaps in response to variation in some other idnetifiable variables.

Although we adopted this strategy for some of our analyses, we still held several reservations about its use. Due to the nature of the MITT design, these lagged analyses necessarily assumed a six-month lag existed in the relationship among the independent and the dependent variables. Not only is the assumption questionable, but the length of the lag makes clear interpretations difficult.

One must assume also that the lagged effects or autocorrelations are representative of what is going on in all the schools at the same time. Thus, we had to assume that the variables themselves changed at about the same time in all schools and that contingencies among variables arise and

disappear at the same time in all schools. An obvious alternative formulation is that the schools do not change in synchrony. Over a particular lag, some schools either may not change at all on certain variables or change in a direction opposite to other schools. A contingency between two variables over a lag in some schools may not appear until some later lag in other schools.

As with the analysis of the impact of the innovation, one may choose to exclude the autoregression term when it correlates substantially with the independent variables of interest. Such a circumstance prevents a clear identification of the effects of the key independent variables because of the large amount of shared variation (Gordon, 1968). Another case for its exclusion is that in which the researcher formally applies a path analytic approach to test a causal model of postulated lagged relationships which specifically excludes the autoregression variable. Including it in such an instance would alter the model under examination.

Controlling for Measurement Error. A more defensible use of the lagged regression analysis is in the control of measurement error. One may wish to examine the relationship between variables as measured at the same point in time but believe that the within-wave correlations are too contaminated with measurement error (due to similarity of instrument format, for example) for an accurate detection of the true relationship by within-wave regression analyses. What suggested this to be the case with many of the MITT variables were suspiciously high within-wave correlations among them.

In order to eliminate the influence of this measurement error, one can examine the relationship between variables across lags. That relationship

can be taken as indicative of the within-wave relationship uncontaminated by the idiosyncracies common to the measurement of both variables at the same time and with the same questionnaire battery (Charters, 1978).

Summary

A first concern with longitudinal analysis was with what kinds of scores to use--the observed scores laden with measurement error or "true" scores from which we excised the error. Of the several methods for estimating true scores, none proved appropriate to the nature of field research data like that in the MITT study. We concluded the observed scores would provide the most interpretable and justifiable data to use in our longitudinal analyses.

The second concern was with what statistical technique to assess change. After considering the merits of repeated measures ANOVA and time-series analysis, we adopted the generalized multiple linear regression approach as versatile and appropriate for field investigations with characteristics of data similar to MITT.

Nevertheless, it has its limitations too. Problems with the impossibility of random assignment to groups can hinder the analysis of treatment/innovation effects. The assumption that the length of a lag between data collections accurately represents the delay in the influence of one variable on another may be highly questionable, particularly given long lags. The most defensible use of the lagged analysis of relationships among variables may be in its use to control for error common to the measurement of variables with the same questionnaire battery at the same point in time.

REFERENCES

Amick, D.J. and Walberg, H.J. Introductory Multivariate Analysis for Educational, Psychological, and Social Research. Berkeley, CA.: McCutchan Publishing Co., 1975.

Cohen, J. and Cohen, P. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: John Wiley & Sons, 1975.

Coleman, J.S. "The Mathematical Study of Change." Methodology in Social Research by H.M. Blalock and A.B. Blalock (eds.). New York: McGraw-Hill, 1968.

Cronbach, L.J. and Furby, L. "How We Should Measure 'Change' -- or Should We?" Psychology Bulletin 74(1), 68-80, 1970. (Also Erratta, Psychology Bulletin 74,218, 1970.)

Hannan, M.T. Aggregation and Disaggregation in Sociology. 1971.

Hannan, M.T. and Young, A.A. Estimation in Panel Models: Results on Pooling Cross Sections in Time Series. Technical Report #51. Stanford University, n.d.

Heise, D.R. "Causan Inference from Panel Data." Sociological Methodology 1970 by E.F. Borgatta and G.W. Bohrnstedt (eds.). San Francisco: Jossey-Bass, 1970.

Jovick, Thomas. "Creating Indices from the Control Structure Interview through Data Collapsing and Multidimensional Scaling: Approaches to Data Analysis in Project MITT." Research report for Project MITT. Eugene,OR.: Center for Educational Policy and Management, University of Oregon, 1978.

Kerlinger, F.N. and Pedhazur, E.J. Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston, Inc., 1973.

Nunnally, J.C. Psychometric Theory. New York: McGraw-Hill, 1967.

O'Connor, E.F., Jr., Response to Cronbach and Furby's "How We Should Measure 'Change' -- or Should We?" Psychology Bulletin 78(2), 159-160, 1972.

_____. "Extending Classical Test Theory to the Measurement of Change." Review of Educational Research 42, 73-97, 1972.

Packard, J.S.; Carlson, R.O.; Charters, W.W., Jr.; Moser, R.H.; and Schmuck, P.A. Governance and Task Interdependence in Schools: First Report of a Longitudinal Study. Eugene, OR.: Center for Educational Policy and Management, University of Oregon, 1976.

Packard, J.S.; Charters, W.W., Jr.; and Carlson, R.O. The Control of Teaching: A Study of Innovating Elementary Schools. Eugene, OR.: Center for Educational Policy and Management, University of Oregon, 1978.

Packard, J.S.; Charters, W.W., Jr.; and Carlson, R.O. Management Implications of Team Teaching: Final Report. Eugene, OR.: Center for Educational Policy and Management, University of Oregon, 1978.

Pelz, D.C. and Lew, R.A. "Heise's Causal Model Applied," in Sociological Methodology 1970 by E.F. Borgatta and G.W. Bohrnstedt (eds.) San Francisco, CA.: Jossey-Bass, 1970.

Wiley, D.E. and Harnischfeger, A. Post Hoc Ergo Proptor Hoc: Problems in the Attribution of Change. Report #7, Studies of Educative Processes. CEMREL Inc., 1973.

_____ and Hornik, R. Measurement Error and the Analysis of Panel Data. Report #7, Studies in Educative Processes. CEMREL Inc., 1973.